DOCUMENT RESUME

ED 294 918 TM 011 650

AUTHOR Linacre, John Michael
TITLE Simple and Effective Algorithms: Computer-Adaptive Testing.
PUB DATE Apr 88
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Algorithms; *Computer Assisted Testing; Cutting Scores; Difficulty Level; Item Banks; Latent Trait Theory

ABSTRACT

Computer-adaptive testing (CAT) allows improved security, greater scoring accuracy, shorter testing periods, quicker availability of results, and reduced guessing and other undesirable test behavior. Simple approaches can be applied by the classroom teacher, or other content specialist, who possesses simple computer equipment and elementary programming skills. Building a CAT system involves techniques required for development of an item bank, selection of items, selection of the items and item order to be presented in a test, and evaluating the test for difficulty. A practical understanding of item response theory and pass-fail criteria is helpful. Although CAT tests are usually intended to determine the examinee's ability level, they can supply diagnostic information as well. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED294918

Simple and Effective Algorithms Computer-Adaptive Testing

John Michael Linacre
University of Chicago
MESA Psychometric Laboratory
5835 S. Kimbark Ave.
Chicago Il 60637
(312) 288-5650

Paper presented at the AERA Annual Meeting,
New Orleans April 6, 1988

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JOHN M. LINACRE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM011650

Computer-adaptive testing

Computer-adaptive testing (CAT) is often presented as complex to understand and difficult to implement. This paper presents three simple approaches which can be applied by the classroom teacher, or other content specialist, who has available simple computer equipment and elementary programming skill.

There are several advantages to CAT:

1) improved security (no test booklets to go astray)
   each examinee has a different test so knowledge of another's test is little help.
2) greater scoring accuracy (no clerical errors)
3) shorter and so faster tests (targeted to examinee)
4) quicker results.
5) reduced guessing and other undesirable test behavior.

These strong points of CAT have been observed by the military (McBride & Martin, 1983). In their application, a CAT test of 15-19 items out of an item bank of 150 items, attained approximately the same reliability and validity. as a conventional test of 30 items. However, this is a rather greater reduction of test length than can usually be expected. Well designed and targeted conventional tests are often only 10-20% longer than the equivalent CAT test. For further discussion of CAT see Weiss and Kingsbury (1984).

CAT testing: What is required

Before a CAT program can be implemented, a pool of test items must be constructed. Each item must be identified with a level of difficulty.

In administering a CAT test, several decisions must be made:

1) What test item to start the test with.
2) How to decide what question to give next.
3) How to decide when to stop the test.

After the test is complete, we should know:
1) What was the ability of the examinee.
2) How accurately do we know that ability.
3) Was the examinee's performance reasonable.

CAT testing: Is a sledge-hammer necessary ?

The CAT literature contains information about a number of techniques that can be applied, for instance:

1) "Maximum information item selection strategy"
"All items in the pool are searched, and the item having the maximum value of information at the individual's current estimated ability is selected for administration" (Weiss 1983).

or

2) "Bayesian item selection strategy"
"an item pool is searched to find the one item that when administered will maximally reduce the posterior variance of an individual's ability estimate" (Weiss 1983).

These techniques, applied exactly as described, have several drawbacks when applied with limited resources. First, they tend to be computation-intensive, difficult to program and slow to execute. Calculating the information or variance corresponding to each item of the bank will involve considerable effort. Obviously it is desirable to short cut this by selecting the next item by a simple technique which is almost as efficient as the optimum method.

However, a question arises about the entire concept of an "optimum method". Choosing the "best" item at every point, particularly with an item pool with an uneven spread of items, often results in many examinees experiencing very similar tests. In fact, if all examinees start with the same "best" item, they will tend to proceed through the CAT test along a restricted number of pathways of "best" items (Stevenson 1987). Consequently, in order to preserve test integrity and avoid dependence on just a few critical items in the pool, it is advisable to select deliberately some less optimum items for administration. Indeed, instead of selecting one optimum item at each point, an item can be selected at random from a short list of near optimum items, and then administered. This will considerably lessen the chance of two examinees experiencing the same test, or of a corrupted item in the bank having more than a minor effect on candidate scores.

Having decided that quick convenient methods are indeed useful, let us examine some of them.

Simple item response theory

The underlying concept behind CAT methods is that each examinee has a certain ability level, and it is the purpose of CAT to estimate it. To enable this requires items with difficulty levels. Then, if the person is more able than the item is difficult, the result is probably a success (or, conversely, a

failure). If a very easy item is administered, nearly every one will succeed and little will be determined about the examinee. Similarly if a very hard item is administered. The simplest strategy for item selection is to find items with difficulties targeted on the estimated ability of the examinee. If the examinee succeeds, we raise his ability estimate, otherwise we lower it.

The basic theory behind this concept is well understood as the Rasch (one-parameter) IRT model:

$$\log (Pni/(1-Pni)) = Bn - Di$$

where Pni = probability of success by examinee n on item i,
and Bn = ability of person n
Di = difficulty of item i.

Over the course of the test, we compare the examinee's probability of success with what actually occurred. After each item, an updated ability estimate is obtained. Finally, after administering a suitable number of items, a good estimate of the examinee's ability will be obtained.

Easy to implement CAT strategies

The use of Rasch IRT methods to select items for administration is well established. Reckase (1974) presented the outline of a program based on the estimation procedure of Wright & Panchepakesan (1969), a procedure now known as "UCON".

A more recent implementation of the same approach is Linacre (1987). This is a program in 172 lines of BASIC which selects and administers multiple-choice test items. On user request, it also recalibrates the item pool. This program has been successfully applied or adapted by a number of practitioners.

CAT from scratch

Building a CAT system from scratch may seem a mind-boggling feat. However:

I. Building the item bank:

On first constructing an item pool in a new content area, there is no need to construct many calibration tests. Rather the content specialist can simply assign each item as it is written to one of a few categories e.g. easy, moderate or hard. For the first round of administrations of the CAT test, the corresponding calibrations could be hard = +1 logit, moderate = 0 logits, easy = -1 logits. At the worst, this design will prevent able examinees wasting time doing easy items, and vice versa for less able examinees.

Save the responses made by each examinee to each item that was administered. The item pool can then be recalibrated, and examinee ability reestimated using standard Rasch software (e.g. MSCALE or MICROSCALE). (Older Rasch programs, such as BICAL, will not be satisfactory because they expect every examinee to have faced every item). Standard software can also be used to obtain other statistics (such as reliability) relating to the entire testing session.

II. Selecting the items:

Bosma (1985) piloted a simple but effective CAT system for implementation by a classroom teacher in the classroom. This used a procedure based on the PROX estimation algorithm (Wright & Stone, 1979). The underlying assumption is that the difficulties of the items actually administered to an individual will be distributed symmetrically around the ability of that individual, which is a reasonable assumption if the item bank is fairly large and the examinee's ability is not out of its range.

In that case, simplifications can be made in estimating the examinee's ability: For the examinee, accumulate the following information:

the number of items administered = L
the number of right answers = R
the number of wrong answers = W
the sum of item difficulties administered = SUMDIF
the sum of squares of item difficulties = SUMSQDIF.

Calculate

the mean item difficulty administered = MEANDIF = SUMDIF/L.

Then the current ability estimate (B) is

B = MEANDIF + OM * LOG (R/W) [LOG = Log to base E]

where

OM = SQuareRoot (1 + (SUMSQDIF/L - MEANDIF**2)/2.89) [**=squared]

This ability estimate is the basis of the difficulty level of the next item to be administered. If the target success rate by the examinee on the items is to be 50%, then the next item to be administered should have difficulty level about equal to B. However, it is often desired to give the examinee a positive experience in an examination so that the examinee leaves the test having a general feeling of success. This can be arranged by administering items on which the examinee has an overall 70% (say) chance of success. That is, using items which have a difficulty about 1 logit less than the examinees estimated ability. This will make for a slightly longer, less efficient, test than optimum, but will still avoid the administration of many inappropriate items.

The appropriate items, from which one is to be selected at random to be administered, are from a certain range or window of item difficulties. A useful window range is one standard error above and below the current ability estimate, or its adjustment for the desired rate of success.

III. Selecting the first item:

Without better information about the examinee, a useful starting rule is to select an item at random from the middle of the pool difficulty range. Until the examinee has at least one item right and one item wrong, his ability is inestimable. To handle this, make the next item about one logit harder (for a string of correct answers) than the previous item until the examinee fails an item. Conversely, one logit easier until the examinee succeeds on an item.

Gan-Ng (1984) based the difficulty of the starting item on the examinee's (or teacher's) response to the question "Are you feeling very smart, smart, or not sure ?".

IV. Stopping the test:

Depending on the purpose for the test, a number of stopping rules are useful:

1) Successful performance:
The test is really too easy for the examinee. The examinee gets nothing wrong. Stop the test after a string of 10 correct answers.

2) Unsuccessful performance:
The test is much too hard for the examinee. The examinee gets nothing right. Unless the test is a minimum competency one, it may be advisable to have the teacher intervene after a string of 5 failed questions, to insure that there is not some major misunderstanding.

3) Usual performance:
The examinee has some successes and some failures. The accuracy with which we know the examinee's ability is measured by the standard error of measurement:

$$SE = OM * SQuareRoot\ (L\ /(R*W))\quad logits$$

This can be calculated after each response, and the test stopped when a suitably small value reached. The relationship between the number of on-target items (close to the final ability estimate) and the standard error is given in the following table:

Number of on-target items	Minimum S.E. of ability estimate
9	.7
11	.6
16	.5
25	.4
45	.3
100	.2

Table 1: On-target test length and minimum S.E. of ability.
(Courtesy of Benjamin Wright)

4) Maximum test length.
An examinee may only miss on, or only succeed on, one or two items. In that case, it may be advisable to set a maximum number of items. This may be determined by exhaustion of likely items from the item pool. (There is no point in presenting items that are obviously much too hard or much too easy).

Testing for a Pass-Fail decision

There are situations in which the intention is not to determine an individuals ability, but just to determine whether it is above or below a certain criterion level.

Wright (1987) suggests the following approach:

Start by administering an item close to the criterion level (C).

Continue by administering items of difficulty D, such that
the number of items administered = L
the number of right answers = R
the number of wrong answers = W
the sum of item difficulties administered = SUMDIF

D = ability estimate B = SUMDIF/L + log(R/W)

Stop when ability estimate is more than 2 standard errors from criterion level:

|B - C| > 2 * SQuareRoot (L / (R*W))

For an examinee with ability 1 logit from the criterion level, this would happen somewhere after 16 on-target questions.

Using CAT tests for diagnosis

CAT tests are usually intended to determine the examinee's ability level. However, they can supply diagnostic information as well.

1) Unexpected successes and failures.

In the course of administering the test, each examinee is likely to meet a few questions that are too hard and a few that are too

easy. If any of these provoke an unexpected response by the examinee they can be reported as misfitting responses for further investigation.

2) Covering the content area.

As a check of the performance an examinee across a subject area, items can be deliberately selected for administration from each part of the area. Again, unexpected successes or failures can be reported.

A simple misfit count can also be calculated to aid the teacher in identifying examinees who perform in an unusual manner:

For each item which is more 'extreme' than one logit from the estimated ability,

$$\text{Misfit \%} = \frac{(\text{unexpected failures} + \text{unexpected successes}) * 100}{(\text{total number of 'extreme' items})}$$

A misfit percent of over 20% would merit closer inspection.

Explaining the CAT test results

As described here, the examinee will be given an ability estimate in terms of "logits" (log odds units). A likely range is -5 to +5. Of course, examinees can still be rank-ordered and their performance reported in terms of percentiles. However, as logits are on a linear scale, a simple transformation may be helpful in reporting these to parents and administrators:

$$\text{Reported score} = (\text{logit} * 10) + 50$$

which will range from 0 to 100.

Together with this score, will be an indication of how much certainty there is about it:

$$\text{Standard error} = (\text{SE} * 10)$$

which will usually range from 3 to 7.

Then a score of +1 logit with standard error of 0.5 logits can be reported as 60 ± 5 points.

Bibliography

Bosma J.F. Teacher and Student Responses to a System for Rational Measurement.
Chicago: University of Chicago. Ph.D. Dissertation. 1985.

Gan-Ng A.-C. The statistical properties of an algorithm for classroom tailored testing.
Chicago: University of Chicago. M.A. Paper 1984

Linacre J.M. A computer program for adaptive testing by microcomputer. Chicago: University of Chicago MESA Memorandum No. 40 1987 (ERIC ED 280 895 Aug. 1987 RIE)

McBride, J.R., Martin, J.T. Reliability and Validity of Adaptive Ability Tests in a military setting. in Weiss D.J. (Ed.) "New Horizons in Testing" New York: Academic Press, 1983

Reckase, M.D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods and Instrumentation 1974 6:2 208-212

Stevenson J. Computerized Adaptive Testing in the Montgomery County, Maryland Public Schools. St. Paul, MN: Assessment Systems Corporation: MicroCAT News April 1987.

Weiss D.J. Introduction in Weiss D.J. (Ed.) "New Horizons in Testing" New York: Academic Press, 1983

Weiss D.J., Kingsbury G.G. Application of computerized adaptive testing to educational problems.
Journal of Educational Measurement 1984 21:4 361-375

Wright, B.D. Testing to a Standard.
Chicago: University of Chicago, MESA 1987.

Wright B.D., Panchepakesan N.A. A procedure for sample-free item analysis.
Educational & Psychological Measurement 1969 29:23-48.

Wright, B.D., Stone, M.H.. Best Test Design.
Chicago, Il: MESA Press. 1979